SCORM 2.0 White Paper:
Stealth Assessment in Virtual Worlds

Valerie J. Shute, Florida State University, vshute@fsu.edu
J. Michael Spector[1], Florida State University, mspector@lsi.fsu.edu

Abstract. This paper envisions SCORM 2.0 as being enhanced by a stealth assessment engine that can be run within games, simulations, and other types of virtual worlds. This engine will collect ongoing and multi-faceted information about the learner while not disrupting attention or flow, and make reasoned inferences about competencies, which form the basis for diagnosis and adaptation. This innovative approach for embedding assessments in immersive virtual worlds (Shute et al., in press) draws on recent advances in assessment design, cognitive science, instructional design, and artificial intelligence (Milrad, Spector & Davidsen, 2003; Shute, Graf, & Hansen, 2005; Spector & Koszalka, 2004). Key elements of the approach include: (a) evidence-centered assessment design, which systematically analyzes the assessment argument, including the claims to be made about the learner and the evidence that supports those claims (Mislevy, Steinberg, & Almond, 2003); (b) formative assessment and feedback to support learning (Black & Wiliam, 1998a; 1998b; Shute, 2008); and (c) instructional prescriptions to deliver tailored content via an adaptive algorithm coupled with the SCORM 2.0 assessments (Shute & Towle, 2003; Shute & Zapata-Rivera, 2008a). Information will be maintained within a student model which provides the basis for deciding when and how to provide personalized content to an individual, and may include cognitive as well as noncognitive information.

Introduction

Measurements are not to provide numbers but insight. Ingrid Bucher

ADL has been criticized for providing a means for achieving accessibility, reusability, interoperability and durability ***only*** for traditional, didactic instruction for individual learners. As early as the 2005 conference ID+SCORM at Brigham Young University, critics challenged ADL to embrace Web 2.0 attributes (e.g., services orientation versus packaged software, an architecture of participation, collective intelligence, data fusion from multiple sources). Clearly the development of self-forming, self–governing online communities of learners has seen far greater uptake than SCORM 1.0.

A significant problem, however, is that the Defense Department – along with many other providers of instruction – simply cannot allow all of its learning to take place in the relatively open fashion common to many Web 2.0 environments. High-stakes programs of instruction leading to certification of competence require a formal process of authentication that recreational learning does not. Currently, Web 2.0 advocates have not been able to recommend a method whereby an instructional program can be authenticated in the absence of authority. Further, there is much evidence that young, inexpert learners often choose exactly that instruction that they do not need (e.g., Clark & Mayer, 2003). DoD must continue to depend on explicitly managed programs of instruction. Games and simulations can, obviously, be used to advantage when they are managed by external authority. Games, simulations, and mission-rehearsal exercises in virtual space also can be used independently by expert learners or teams of learners who have the capacity to determine when intended outcomes are, or are not, realized. There are limits. That may be about to change.

Proposed. We propose that SCORM 2.0 should expand on an innovative approach for embedding assessments in immersive games (Shute et al., in press), drawing on recent advances in assessment design, cognitive science, instructional design, and artificial intelligence (Milrad, Spector & Davidsen, 2003; Shute, Graf, & Hansen, 2005; Spector & Koszalka, 2004). Key elements of the approach include: (a) evidence-centered assessment design, which systematically analyzes the

[1] After Jan. 5, 2009, Dr. Spector will be at the University of Georgia, Learning and Performance Laboratory, Athens, GA.

assessment argument, including the claims to be made about the learner and the evidence that supports (or fails to support) those claims (Mislevy, Steinberg, & Almond, 2003); (b) formative assessment to guide instructional experiences (Black & Wiliam, 1998a; 1998b); and (c) instructional prescriptions to deliver tailored content via two types of adaptation: micro-adaptation and macro-adaptation, addressing the *what to teach* and *how to teach* parts of the curriculum (Shute & Zapata-Rivera, 2008a).

To accomplish these goals, an adaptive system will comprise the foundation for SCORM 2.0 assessments (Shute & Towle, 2003; Shute & Zapata-Rivera, 2008a). This will enable an instructional game/ simulation/ virtual world to adjust itself to suit particular learner/player characteristics, needs, and preferences. Information will be maintained within a student model, which is a representation of the learner (in relation to knowledge, skills, understanding, and other personal attributes) managed by the adaptive system. Student models provide the basis for deciding when and how to provide personalized content to a particular individual, and may include cognitive as well as noncognitive information.

Research/Technology to Support SCORM 2.0

We cannot direct the wind but we can adjust the sails. Anonymous

Assumptions. The main assumptions underlying this white paper are that: (a) learning by doing (required in game play) improves learning processes and outcomes, (b) different types of learning and learner attributes may be verified and measured during game play, (c) strengths and weaknesses of the learner may be capitalized on and bolstered, respectively to improve learning, and (d) formative feedback can be used to further support student learning (Dewey, 1938; Gee, 2003; Shute, 2007; Shute, 2008; Shute, Hansen, & Almond, 2007; Squire, 2006). In addition, these assumptions represent the legitimate reasons why DoD components and its vendors seek to exploit the instructional affordances of games.

The Idea. New directions in psychometrics allow more accurate estimations of learners' competencies. New technologies permit us to administer formative assessments during the learning process, extract ongoing, multi-faceted information from a learner, and react in immediate and helpful ways. This is important given large individual differences among learners, and reflects the use of adaptive technologies described above. When embedded assessments are seamlessly woven into the fabric of the learning environment so that they are virtually invisible or unnoticed by the learner, this is *stealth assessment*. Stealth assessment can be accomplished via automated scoring and machine-based reasoning techniques to infer things that would be too hard for humans (e.g., estimating values of evidence-based competencies across a network of skills). A key issue involves not the collection or analysis of the data, but making sense of what can potentially become a deluge of information. This sense-making part of the story represents a complementary approach to the research proposed by Rachel Ellaway.

Another major question concerns the best way to communicate learner-performance information in a way that can be used to easily inform instruction and enhance learning. Our solution to the issue of making sense of data and fostering learning within gaming environments is to extend and apply evidence-centered design (ECD; Mislevy, Steinberg, & Almond, 2003). This provides a way of reasoning about assessment design, and a way of reasoning about learner performance in a complex learning environment, such as an immersive game.

The Methodology. There are several problems that must be overcome to incorporate assessment in games. Bauer, Williamson, Mislevy, and Behrens (2003) address many of these same issues with respect to incorporating assessment within interactive simulations in general. In playing games, learner-players naturally produce rich sequences of actions while performing complex tasks, drawing upon the very skills we want to assess (e.g., communication skill, decision making, problem solving).

Evidence needed to assess the skills is thus provided by the players' interactions with the game itself – the processes of play – which may be contrasted with the product(s) of an activity, as is the norm within educational, industrial, and military training environments. Making use of this stream of evidence to assess knowledge, skills, and understanding presents problems for traditional measurement models used in assessment. First, in traditional tests the answer to each question is seen as an independent data point. In contrast, the individual actions within a sequence of interactions in a simulation or game are often highly dependent on one another. For instance, what one does in a flight simulator or combat game at one point in time affects subsequent actions later on. Second, in traditional tests, questions are often designed to get at one particular piece of knowledge. Answering the question correctly is evidence that one knows a certain fact; i.e. one question – one fact.

By analyzing responses to all of the questions or actions taken within a game (where each response/action provides incremental evidence about the current mastery of a specific fact, concept, or skill), instructional or training environments may infer what learners are likely to know and not know overall. Because we typically want to assess a whole constellation of skills and abilities from evidence coming from learners' interactions within a game or simulation, methods for analyzing the sequence of behaviors to infer these abilities are not as obvious. ECD is a method that can address these problems and enable the development of robust and valid simulation- or game-based learning systems. Bayesian networks comprise a powerful tool to accomplish these goals. ECD and Bayes nets will each be described in turn.

Evidence-centered design. A game that includes stealth assessment must elicit behavior that bears evidence about key skills and knowledge, and it must additionally provide principled interpretations of that evidence in terms that suit the purpose of the assessment. Figure 1 shows the basic models of an evidence-centered approach to assessment design (Mislevy, Steinberg, & Almond, 2003).

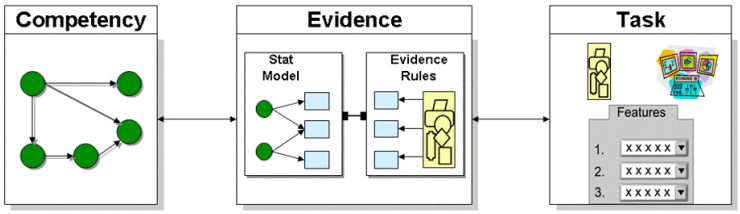

Figure 1. The Central Models of an Evidence-Centered Assessment Design

Working out these variables and models and their interrelationships is a way to answer a series of questions posed by Messick (1994) that get at the very heart of assessment design:

- ***What collection of knowledge and skills should be assessed?*** (Competency Model; CM). A given assessment is meant to support inferences for some purpose, such as grading, certification, diagnosis, guidance for further instruction, etc. Variables in the CM are usually called 'nodes' and describe the set of knowledge and skills on which inferences are to be based. The term 'student model' is used to denote a student-instantiated version of the CM—like a profile or report card, only at a more refined grain size. Values in the student model express the assessor's current belief about a learner's level on each variable within the CM.
- ***What behaviors or performances should reveal those constructs?*** (Evidence Model; EM). An EM expresses how the learner's interactions with, and responses to a given problem constitute evidence about competency model variables. The EM attempts to answer two questions: (a) What behaviors or performances reveal targeted competencies? and (b) What is the connection between those behaviors and the CM variable(s)? Basically, an evidence model lays out the argument about why and how the observations in a given task situation (i.e., learner

performance data) constitute evidence about CM variables. This comprises the statistical *glue* in the ECD approach.

- ***What tasks should elicit those behaviors that comprise the evidence?*** (Task Model; TM). TM variables describe features of situations (e.g., scenarios) that will be used to elicit performance. A TM provides a framework for characterizing and constructing situations with which a learner will interact to provide evidence about targeted aspects of knowledge related to competencies. These situations are described in terms of: (a) the presentation format, (b) the specific work or response products, and (c) other variables used to describe key features of tasks (e.g., knowledge type, difficulty level). Thus, task specifications establish what the learner will be asked to do, what kinds of responses are permitted, what types of formats are available, and other considerations, such as whether the learner will be timed, allowed to use tools (e.g., calculators, the Internet), and so forth. Multiple task models can be employed in a given assessment. Tasks (e.g., quests and missions, in games) are the most obvious part of an assessment, and their main purpose is to elicit evidence (directly observable) about competencies (not directly observable).

In games with stealth assessment, the student model will accumulate and represent belief about the targeted aspects of skill, expressed as probability distributions for student-model variables (Almond & Mislevy, 1999). Evidence models would identify what the student says or does that can provide evidence about those skills (Steinberg & Gitomer, 1996) and express in a psychometric model how the evidence depends on the competency-model variables (Mislevy, 1994). Task models would express situations that can evoke required evidence. The primary tool to be used for the modeling efforts will be Bayesian networks.

Bayesian networks. Bayesian networks (Pearl, 1988) are used within student models to handle uncertainty by using probabilistic inference to update and improve belief values (e.g., regarding learner competencies). The inductive and deductive reasoning capabilities of Bayesian nets support "what-if" scenarios by activating and observing evidence that describes a particular case or situation, and then propagating that information through the network using the internal probability distributions that govern the behavior of the Bayesian net. Resulting probabilities inform decision making, as needed in, for instance, the selection of the best chunk of training support to subsequently deliver based on the learner's current state. Examples of Bayes net implementations for student models may be seen in: Conati, Gertner, and VanLehn (2002), Shute, Graf, and Hansen (2005); and VanLehn et al. (2005).

Bayesian Example. To illustrate how this proposed methodology will actually work inside of a game, we have implemented a "toy" model using a Bayesian network approach.[2] Imagine that you—in the game and in real life—are a newly deployed young officer in a combat zone, and you find yourself in charge while the CO and/or XO are away from the command post. It is 1700 hrs when the radio and FAX lines begin buzzing. Information on two new skirmishes and suspicious activity (10 KM away) comes in requiring your attention. You answer the radio, read the fax, and start posting information to maps, and filling in journal entries. The radio and FAX continue to buzz—disrupting your transcription. But you choose to ignore them and continue writing in the journal (although you make some inaccurate entries given the distractions). Meanwhile, the stealth assessment in the game is making inferences about how well you are performing. The model in Figure 2 shows the state of the model with no information of you at all, just the prior and conditional probabilities.

[2] Note: We have made up a few example competencies for this illustration.

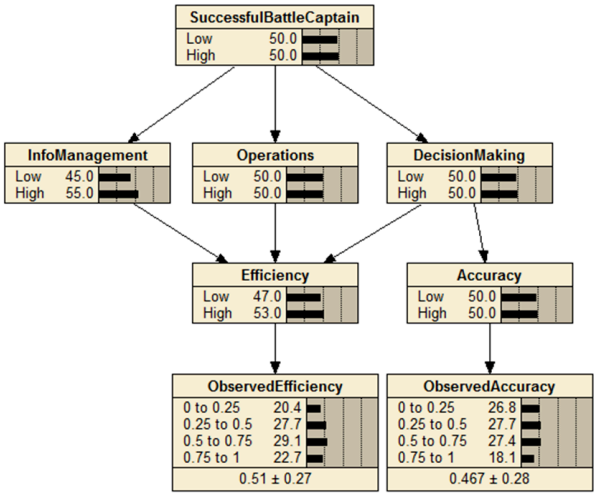

Figure 2. Hypothetical Bayes model to illustrate command performance

Now look at the right side of the model—the "Decision Making" branch. Given your observed performance (i.e., inaccurate notes and inefficient response to incoming communications), the system would update the model to yield something like Figure 3. Clearly, your decision making skills need to be improved.

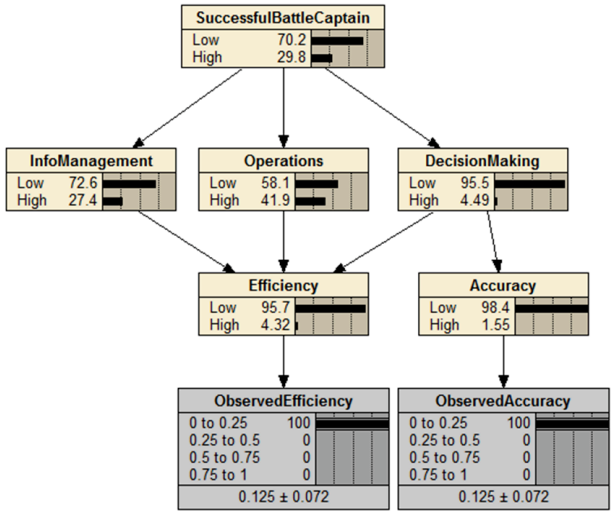

Figure 3. Hypothetical Bayes model showing marginal probabilities after observing low efficiency and low accuracy actions.

Actions can be captured in real-time as the learner interacts with the game, and associated indicators can be used to provide evidence for the appropriate competencies. This would be accomplished via the evidence model. Each node has two or more discrete states (e.g., low and high). Marginal probabilities are presented for each state. The lower two evidence-model nodes (shaded grey) represent continuous variables, discretized into four states, ranging from 0 to 1, that can be used to model specific actions.

Systems Thinking. In addition to the ideas about stealth assessment (above), another research stream we recommend incorporating into SCORM 2.0 relates to the externalization of internal representations (i.e., mental models). While internal representations are not available for direct and immediate observation (similar to unobservable latent traits, like decision-making ability in the prior

example); the quality of internal representations is closely associated with the quality of learning. So, to help improve support for learning/understanding, we have devised a theoretical foundation and a collection of tools to facilitate assessment that can (a) externalize these internal mental models or current *belief nets*; and (b) provide personalized, reflective, and meaningful feedback to learners, particularly in relation to complex, challenging problem-solving domains (Shute, Jeong, & Zapata-Rivera, in press; Shute, Jeong, Spector, Seel, & Johnson, in press). Working within such ill-structured and dynamic domains (manifest in rapidly-changing problem scenarios) is exactly what is encountered by many military and other jobs, especially under conditions of rapidly-changing events. Assessing how the mental models change relative to an expert's representation will provide the basis for useful interventions.

Our approach to representing a learner's (or group's) current set of beliefs about a topic is to overlay Bayesian networks on concept maps (or causal diagrams). This permits us to model and question the degree to which relationships among concepts/nodes hold as well as the strength of the relationships. In addition, prior probabilities can be used to represent preconceived beliefs. This probabilistic network provides a rich set of modeling tools that we can use to represent the degree to which individuals (or teams) ascribe to a particular belief pattern. Incorporating an assessment layer on top of the concept maps lets us flesh out the maps more fully, resulting in a collection of evidence from learners in terms of their evolving mental models as indicated by their relationship to the strength and relevance of associations, directionality of the stated relations, and the specified type or nature of the relationship. The result is a set of flexible belief networks (or FBNs) (Shute & Zapata-Rivera, 2008b).

Conclusion. Given the overwhelming complexity and quantity of data that can be produced from close examination of performance data, concept maps, and cognitive processes at the granular level, new software tools and methods we're developing can produce valid inferences of competencies, as well as visual representations that can simultaneously reveal: (a) global patterns emerging in the performance data and maps, as well as the cognitive processes, events, and/or conditions that trigger changes; (b) the extent to which the changing patterns are progressing toward a target model; and (c) detailed and precise information on what and where changes are occurring within the maps. Together, the two research streams (stealth assessment and modeling systems thinking) will support the assessment and diagnosis of specific knowledge, skills, and understanding. Moreover, the research will provide clear prescriptions and standard implementation protocols for valuable training to improve performance in relation to important and complex jobs.

We end with a series of questions and answers relating to some anticipated issues.

Questions/Answers

1. *How will the proposed work break new ground in the area of adaptive training technology? What will the technology be able to accomplish that could not be accomplished before? How will the work performed expand on established capabilities?*

Innovations 1 & 2. The goal of an adaptive training system is to create an instructionally sound and flexible environment that supports knowledge and skill acquisition for individuals with a range of abilities, disabilities, interests, backgrounds, and other characteristics (Shute & Zapata-Rivera, 2008a). Current adaptive training systems vary in many ways and are constructed in an eclectic and generally ad hoc manner. While some of these adaptive training systems have proven effective, there is a general lack of a systematic instructional design approach for adaptive systems, and this is especially true with regard to game-based training in complex domains for adults. Our idea is to adopt a mastery approach to adaptive, game-based training. Older training systems often hold instructional time constant and allow achievement to vary. While this is efficient in terms of training time, it often fails to ensure competency development for a wide range of trainees. Because individuals differ so much in terms of incoming knowledge and skills as well as learning speed, it is possible to hold achievement constant and allow or require most learners to achieve mastery of the

training objectives – this is called a mastery approach. While mastery learning has been tried in non-game-based contexts, it has not been emphasized in game-based learning environments. The main characteristics of such a mastery learning approach are: (a) the establishment of a criterion level of performance held to represent mastery of a given skill or concept, (b) frequent assessment of student progress toward the mastery criterion, and (c) provision of formative feedback (Shute, 2008) to enable learners who do not initially meet the mastery criterion to do so on later assessments. This idea uses a mastery focus to improve learning effectiveness. Moreover, the entire training experience can be made engaging and compelling by embedding tasks/missions within an immersive game. As a result, we can envision learners clamoring to play (even in their off-time), boasting of their current level/rank in the game, and thus yielding a healthy supply of potential workers who are certifiably competent in terms of the requisite knowledge and skills. Our mastery approach and the associated gaming environment represent the first two areas of innovation.

Innovation 3. The primary challenge of accomplishing the goals of any successful adaptive system depends largely and squarely on accurately identifying characteristics of a particular learner or group of learners—such as type and level of current knowledge, skills, and understanding—and then determining how to leverage the information to improve subsequent learning (e.g., Conati, 2002; Park & Lee, 2008; Shute, Lajoie, & Gluck, 2000). This occurs during the learning process (e.g., during game play) and can be thought of as a set of small, ongoing, stealth assessments (noted above as an important part of the mastery approach). Decisions about content selection are typically based on performance and on subsequent inferences of students' knowledge and skill states. The success of any adaptation is a function of the validity and reliability of the underlying assessments. Consequently, this innovation relates to the use and extension of evidence-centered assessment design (Mislevy, Almond, & Steinberg, 2003) within a gaming environment. That is, we can expand current capabilities relating to our stealth assessment approach (Shute et al., in press) that would ultimately be embedded within an intelligent, immersive game environment to assess, diagnose, and enhance knowledge, skills, and understanding.

Innovation 4. Finally, another area that will be advanced concerns our planned use of both criterion- and norm-referenced measures within the given game. Criterion-referenced testing (CRT) is based on a well-specified domain, which we'll accomplish via our ECD formulation (i.e., the competency model). This yields items that are appropriately crafted to elicit evidence that will provide information to the competency model (which in turn will enable inferences about the degree of mastery a learner attains in relation to the particular knowledge and skills in the model). We can accomplish these evidence-based, valid inferences via the Bayes net that operates within the student model. Scores on criterion-referenced assessments will simply (and stealthily) be collected during the course of game play and indicate what individuals can do, and how well they can do it. Using CRT is intended to support learners' intrinsic motivation—to improve "personal best" scores. On the other hand, norm-referenced testing (NRT) compares a person's score against the scores of a group of people who have already played the game. This can invoke a competitive environment, which may serve as a source of extrinsic motivation. NRTs are designed to rank-order individuals in order to compare scores, which is an integral feature of games. Combining CRT and NRT under one "game roof" represents our fourth innovation.

Potential benefits relating to the ideas and technologies described in this white paper include the scalable design of an effective, efficient, and enjoyable training/learning experience for individuals. This innovative paradigm is also intended to improve the validity of assessment and diagnosis as well as the efficacy of the instructional intervention by providing a common evidence-based structure, language, and technology, founded on sound probability theory.

2. *Will the models that are developed to support the functioning of the training technology be pre-scripted? Will they be run-time? Will they be predictive? Will they be static or dynamic? How will these models deal with ill-defined aspects certain jobs?*

Several kinds of models may be included in the SCORM 2.0 assessment engine (e.g., CM, EM, TM). The question of whether the models are canned or created on-the-fly is interesting because it highlights another aspect of our innovative approach. Some aspects of the models are pre-constructed (e.g., specification and structure of the competencies to be assessed and bolstered), but the estimates of learner mastery will constantly evolve over time. That is, mastery estimates will change based on the actions of an individual player, represented within the dynamic student model. The heart of the system is the notion of a dynamic learner profile that is constantly being updated based on individual actions and decisions in the game. In addition, as more players engage in the game, the entire competency model will be progressively updated and refined, with respect to the prior and conditional probabilities. Thus the competency model "learns" and becomes more accurate over time. These capabilities are part of the nature of Bayes nets. Our rationale for using Bayes nets is because we are not able to observe what learners know and can do directly, so we gather evidence in the form of responses to relevant tasks. But tasks are not perfect and people do not respond with perfect consistency; thus we need to express claims with some degree of uncertainty, and Bayes nets are perfect for that. In addition, Bayes nets represent a very powerful tool for allowing probabilistic inferences: (a) *diagnostic inferences* (from effects to causes—e.g., what is this learner's current area(s) of difficulty?), and (b) *predictive inferences* (from causes to effects—e.g., given this student's performance history, how soon is she likely to be fully competent?).

The models will be able to handle ill-defined aspects of certain jobs. As illustrated in Shute et al. (in press), slippery, ill-defined attributes such as "creative problem solving" can be decomposed into constituent knowledge and skills to the point where we can specify reliable indicators of performance, comprising the observable, measurable data. Once the competency model has been established (with help from SMEs), we will be able specify the observables, and associate relevant scoring rules.

3. *How will learning experiences be selected and composed during training? What will determine the mix of experiential vs. didactic content? How will lesson content be constructed? What are the benefits and risks of different potential approaches?*

First, the notion of 'lesson' in our game-based approach is not at all like a traditional lesson. Instead, the approach involves the design of scenarios (as brief as 5 minutes or as long as several hours) that include a number of activities that can vary in terms of difficulty level. Each activity and the associated tasks in a given scenario will be mathematically and explicitly connected (via an evidence model) to the appropriate competency or competencies. Because our approach involves an immersive, adaptive game, the content will appear to be different for each player (i.e., a function of the player's performance history), but ultimately connected to the competency model.

Because we suggest using task models as part of our ECD approach to assessment design, we can generate scenarios dynamically. The content of each scenario will be specified, generally, in advance. However, task models will permit the system to instantiate variables within a particular scenario shell to produce variants on the content. For example, if a player succeeded in solving a relatively easy task, the next task the learner encounters would be more difficult.

The learning experiences will be embedded within scenarios and consist of particular tasks or problems to solve. Thus a learning experience will be selected on the basis of the current state of the student model. For example, competencies in the model that are indicated as almost attained (comparable to Vygotsky's zone of proximal development) will comprise the pool of candidate competencies for subsequent assessment and instruction. Each competency has its own set of linked scenarios. Typical adaptive training systems may use an instructional approach that selects content based on the learner's request, or based on an algorithm that chooses a learning object to be instructed or practiced. These learning objects are often presented as standalone content, with insufficient context. We plan to take a different approach—one that is aligned with van Merriënboer's 4C/ID

model (Four-Component Instructional Design Model; van Merriënboer & Kirshner, 2007). That is, within the game environment, learners will engage in solving *integrated* problems and whole tasks in each scenario, designed so that constituent skills develop in a rich and coordinated manner. In addition, tasks will be connected to the broader competencies (multiple skills, not single ones), feedback (implicit and explicit) would continually support their learning, and just-in-time information and part-task to whole-task practice would foster the integration of knowledge and skills, making whole-task practice representative of their real-life task performance.

Regarding the mix of experiential vs. didactic content, our game-based approach will have an experiential feel. Actions taken within a scenario update the student model and inform subsequent "content" (i.e., scenarios). Information that is usually made explicit in didactic instruction will be embedded in each scenario, so that learners have to understand the basic information in order to master the skills associated with that scenario. Players will also be able to obtain assistance from "embedded experts" if they get stuck. In that case, the player could freeze frame the game, and click on the picture/avatar of an "expert" to get the expert's model or solution to a particular problem. While this does not didactically tell the player what to do and what the basic information is, it provides information required to succeed and it also provides valuable information about what other, more skillful players or experts have done in similar situations (like worked examples). We also would like to encourage collaboration among players to promote learning. For instance, we would encourage players to post and respond to other posts on various strategies on a discussion forum. Such exchange of strategies occurs naturally in gaming cultures.

Another part of the game will include providing medals to players who achieve mastery of different and important competencies (the equivalent of merit badges in the scouts). Furthermore, there would be one mega-medal for achieving, for instance 90% mastery on the highest node in the competency model. This would require many hours of game play to achieve, analogous to beating the highest level in typical immersive games.

Personnel. Dr. Shute is an associate professor at Florida State University (Instructional Systems program). She recently (Aug. 2007) came to FSU from Educational Testing Service where she designed and developed basic and applied research projects related to assessment, cognitive diagnosis, and learning from advanced instructional/assessment systems. Before coming to ETS in 2001, Dr. Shute worked in industry for two years—again in the area of advanced instructional system design and development. Prior to that, she was employed at the Air Force Research Lab, Brooks Air Force Base, Texas for 13 years. Dr. Shute's primary research interests include basic and applied research examining cognitive process measures, as well as designing, developing, and evaluating advanced systems to support learning. She has written more than 100 journal articles, book chapters, and technical papers; co-edited two books. Dr. Spector is the Associate Director of the Learning Systems Institute at FSU and has been conducting research in assessing learning in complex domains for more than 10 years. Dr. Spector was the senior scientist for instructional systems research at the Air Force Research Lab and has published six books and more than 75 journal articles in areas related to this proposed effort. In addition, he is a member of the IEEE Learning Technology Technical Committee working on curricula for advanced learning technology along with Kinshuk, Hartley, Koper et al.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 223-237.

Bauer, M., Williamson, D., Mislevy, R. & Behrens, J. (2003). Using Evidence-Centered Design to develop advanced simulation-based assessment and training. In G. Richards (Ed.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003* (pp. 1495-1502). Chesapeake, VA: AACE.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: School of Education, King's College.

Clark, R. C., & Mayer, R. E. (2003). *e-Learning and the science of instruction*: Pfeiffer, San Francisco.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence, 16*(7/8), 555–575.

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *Journal of User Modeling and User-Adapted Interaction, 12*(4), 371-417.

Dewey, J. (1938). *Experience and education*. New York: Simon & Schuster.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Education Researcher, 32*(2), 13–23.

Milrad, M., Spector, J. M., & Davidsen, P. I. (2003). Model Facilitated Learning. In S. Naidu (Ed.), *Learning and teaching with technology: Principles and practices* (pp. 13-27). London: Kogan Page.

Mislevy, R. J., (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Park, O., & Lee, J. (2008). Adaptive Instructional Systems. Technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology* (3rd ed.). Mahwah, NJ: Erlbaum Associates.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Kaufmann.

Shute, V. J. (2007). Tensions, trends, tools, and technologies: Time for an educational sea change. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 139-187). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Shute, V. J. & Towle, B. (2003). Adaptive e-learning. *Educational Psychologist, 38*(2), 105-114.

Shute, V. J. & Zapata-Rivera, D. (2008a). Adaptive technologies. In J. M. Spector, D. Merrill, J. van Merriënboer, & M. Driscoll (Eds.), *Handbook of Research on Educational Communications and Technology (3rd Edition)* (pp. 277-294). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J. & Zapata-Rivera, D. (2008b). Using an evidence-based approach to assess mental models. In D. Ifenthaler, P. Pirnay-Dummer, & J. M. Spector (Eds.), *Understanding models for learning and instruction: Essays in honor of Norbert M. Seel*. New York: Springer.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for sighted and visually-disabled students. In L. PytlikZillig, R. Bruning, & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169–202). Greenwich, CT: Information Age Publishing.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: Designing for learning effectiveness and accessibility*, ETS Research Report, RR-07-26 (pp. 1-54), Princeton, NJ.

Shute, V. J., Lajoie, S. P., & Gluck, K. A. (2000). Individualized and group approaches to training. In S. Tobias & J. D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military* (pp. 171–207). New York: Macmillan.

Shute, V. J., & Jeong, A., C., & Zapata-Rivera, D. (in press). Using flexible belief networks to assess mental models. In *Instructional Design for Complex Learning*. New York, NY: Springer.

Shute, V. J., Jeong, A. C., Spector, J. M., Seel, N. M., & Johnson, T. E. (in press). Model-based methods for assessment, learning, and instruction: Innovative educational technology at Florida State University. To appear in M. Orey (Ed.), *2009 Educational Media and Technology Yearbook*, Westport, CT: Greenwood Publishing Group.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (in press). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *The Social Science of Serious Games: Theories and Applications*. Philadelphia, PA: Routledge/LEA.

Spector, J. M., & Koszalka, T. A. (2004). *The DEEP methodology for assessing learning in complex domains* (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building). Syracuse, NY: Syracuse University**.**

Squire, K. D. (2006). From content to context: Videogames as designed experience. *Educational Researcher, 35*(8), 19-29.

Steinberg, L. S., & Gitomer, D. G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science, 24*, 223-258.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence and Education, 15*(3), 147-204.

van Merriënboer, J.J.G & Kirshner, P. (2007). *Ten steps to complex learning*. Mahwah, NJ: Erlbaum.